Anim. Behav., 1992, **44**, 833–841

The chi-squared statistic in ethology: use and misuse

MATTHEW KRAMER* & JAMES SCHMIDHAMMER
Department of Statistics, University of Tennessee, Knoxville, TN 37996, U.S.A.

(Received 1 April 1991; initial acceptance 17 May 1991; final acceptance 29 January 1992; MS. number: A6017)

Abstract. Pearson's chi-squared and related tests are not appropriate for all frequency-type data. Lack of independence between observations can invalidate traditional contingency table analysis because sampling distributions are no longer Poisson, multinomial or product multinominal. The usual consequence is that a true null hypothesis is rejected too often, making dubious a claim of significance. If possible, counts should be verified as coming from a Poisson or multinomial distribution before conducting tests. Assuming independence is not sufficient; chi-squared and related tests are shown not to be robust to the violation of this assumption. Frequency-type ethological data, such as the number of encounters between individuals or performances of a behaviour, are likely to violate the assumption of independence. A superior approach for the analysis of these data is demonstrated using parametric and non-parametric analysis of variance (ANOVA).

After collecting frequency-type data, many researchers turn to the well-known Pearson's chi-squared statistic for tests of significance. Here we discuss a common problem that arises from the improper use of chi-squared and related tests (e.g. the likelihood ratio G^2 test or Fisher's exact test) of frequency-type data. We illustrate these problems with simulated ethological data. The problem stems from an inadequate understanding of when a chi-squared or related test is appropriate. Statistical texts directed at biologists, for example, Sokal & Rohlf (1981), compound this problem by giving insufficient or enigmatic treatment of the assumptions underlying chi-squared tests. Many researchers believe chi-squared and related tests to be 'distribution free' or 'non-parametric'. This is likely due to their relegation to the non-parametric statistics section in many textbooks or discussions of them in books on non-parametric statistics. As will be explained below, these tests are not 'distribution free'.

The greatest problem arises from lack of independence between observations. This problem also exists in the psychological literature, noted long ago by Lewis & Burke (1949). Lack of independence among observations can lead to an inflated chi-squared statistic (statistical significance is more likely to be declared incorrectly) or deflated chi-squared statistic (the test is too conservative).

Pearson's chi-squared and related tests are traditionally used (1) to test for the goodness-of-fit of an empirical distribution to a theoretical distribution, and (2) to test for independence of two or more variables. In a goodness-of-fit test, one expects a certain number of observations to fall into each of the classes based on theory, for example, Mendelian genetic ratios. If the observed and expected frequencies are similar to each other in each of the classes, then the fit is good. The observations may be discrete, for example, counts of individuals, or continuous, for example, the duration of a bird song. Since chi-squared and related tests are based on the number of observations in a class, information is lost if the observations are continuous. In tests of independence, the observations are cross-classified using two or more classifying variables. One wants to assess whether the variables used for cross-classifying are independent of each other. The problem of dependence between observations affects both tests of independence and goodness-of-fit in the same way.

SAMPLING DISTRIBUTIONS

Traditional contingency tables are generally tested for independence among the classifying variables

*Present address: Statistical Research Division, U.S. Bureau of the Census, Room 3000-4, Washington, DC 20233, U.S.A.

0003–3472/92/110833+09 $08.00/0

using a chi-squared or G^2 statistic. These tests are valid only under certain sampling distributions. The three most commonly encountered distributions are (1) independent Poisson sampling, (2) simple multinomial sampling, and (3) product multinomial sampling.

In (1) independent Poisson sampling, each cell count is a sample from a Poisson distribution. The mean and variance of this distribution are equal to each other, an important feature that will be discussed in greater detail below. There is no restriction on total sample size in independent Poisson sampling. An example of data that might be collected under this sampling scheme is the following. A researcher counts male frogs along the lake shore until a certain portion of the lake shore has been examined. Each frog is classified by two criteria, whether it is vocalizing and whether it is in the water. One must first verify that the presence or absence of vocalization or location of one frog does not influence these qualities in its neighbours, since this would violate the assumption of independent observations.

In (2) simple multinomial sampling, each cell count is assumed to come from a multinomial distribution. A restriction has been placed on the data since the sample size is predetermined. One may think of this distribution as independent Poisson sampling occurring in each cell with a restriction on the contingency table. With this restriction, the cell frequencies of a contingency table built from samples from a multinomial distribution must add up to N, the previously determined sample size. An example of simple multinomial sampling is the following. The ethologist researching frogs above decides to stop after categorizing 100 frogs.

In (3) product multinomial sampling, an additional restriction is placed on one or more of the marginal totals. For example, if sex were one factor, a researcher might collect data on 50 individuals of each sex. In these three sampling distributions the likelihood functions yielding the expected cell frequencies are usually identical. Thus, for practical purposes, statistical testing is equivalent (Bishop et al. 1975). In general, sampling from one of the above distributions has occurred when the count of each cell in a contingency table represents the number of independent events, observations or individuals possessing the combination of characteristics identifying that cell (e.g. vocalizing male frogs out of water). One may then proceed with a traditional contingency table analysis.

Much of the published theoretical work implicitly assumes that sampling has occurred from one of the above distributions. Thus, it may not be surprising that this assumption is rarely emphasized in textbooks. This may have lulled some researchers into believing that hypotheses concerning association for any discrete data arrayed in a contingency table are appropriately tested using a chi-squared or related statistic.

VIOLATING INDEPENDENCE

The term 'independence' may be confusing. Independence can refer to independence among variables (often the stated null hypothesis in contingency table analysis) or to independence among the observations that fill contingency tables. One may have independent observations but dependence among variables, or dependence among observations but independence among variables. In traditional contingency table analysis, the observations that are summed to yield the cell frequencies must be independent of each other regardless of which of the three sampling distributions discussed above is used. If the observations are not independent, they cannot come from one of the three distributions listed above and traditional contingency table analysis is incorrect. A violation of this assumption results in altered chi-squared values. We emphasize the relationship between independence of observations and the Poisson distribution. If the observations are independent of each other in time or space, each cell contains a sample from a Poisson distribution. If a cell does not contain a sample from a Poisson distribution, observations are not independent and traditional contingency table analysis can not be used. Thus, chi-squared and related tests are not 'distribution free', since cells must contain samples from a Poisson distribution.

Chi-squared Test

To help understand the relationship between the Poisson distribution and the chi-squared test, it may be useful to review the philosophy behind the chi-squared test. Rigorous mathematical explanations of chi-squared and related tests are available elsewhere in highly readable form (e.g. Cochran 1952; Bishop et al. 1975; Moore 1986). The chi-squared test is based on the χ^2 distribution. The χ^2

distribution is the sum of squared standardized normal variables. Each normal variable is standardized (denoted by z) by subtracting its mean and then dividing by its standard deviation,

$$z=\frac{\bar{X}-\mu}{\sigma}.$$

A χ^2 distribution with 1 degree of freedom is the distribution of one squared standardized normal variable, i.e.

$$\frac{(\bar{X}-\mu)^2}{\sigma^2}.$$

A χ^2 distribution with two degrees of freedom is the distribution of the sum of two independent ones, i.e.

$$\sum_{i=1}^{2}\frac{(\bar{X}_i-\mu_i)^2}{{\sigma_i}^2},$$

and so forth. Note the similarity to the chi-squared test statistic,

$$\sum\frac{(\text{observed}-\text{expected})^2}{\text{expected}},$$

if we say the expected frequency in the numerator is the mean, μ_i, and the expected frequency in the denominator is the variance, ${\sigma_i}^2$. In the Poisson distribution the mean and variance are equal. Since the observed frequency in each cell is assumed to be a sample from a Poisson distribution, the mean (= variance) is estimated by the expected frequency. As the expected frequency increases, Poisson variables behave more and more like normal variables. Thus, we should expect to get roughly a χ^2 distribution with k degrees of freedom by summing k independent Poisson variables, after transforming each of them using the chi-squared test statistic algorithm; subtracting the mean and squaring the result, then dividing by the mean (= variance). As in other statistical tests, degrees of freedom are lost when parameters (here, cell means) are estimated from the data.

Dependence Between Behaviour Patterns

Behaviour patterns are rarely, if ever, performed independently of each other (imagine the efficiency of an animal whose behaviour patterns occur at random in time rather than in response to internal and external conditions). Many patterns of behaviour that interest ethologists tend to occur in clusters. This has important statistical consequences: (1) individual behaviour patterns are not independent, and (2) the variance of the probability distribution of the behaviour is inflated (Gleser & Moore 1983).

Suppose a behaviour tends to be performed in clusters. Furthermore, suppose that the mean frequency of this behaviour does not change from one time period to another. In each time period, the clustering would cause the data to be overdispersed relative to a Poisson distribution, i.e. the variance would be larger than the mean. Since the variance is larger, the frequencies between time periods would tend to differ more than they would if they were independent Poisson samples. When constructing the chi-squared statistic, this would not be taken into account. The consequence is an inflated chi-squared statistic. The same problem arises in tests of independence for the same reason. Below we give results, using simulated data from a variety of discrete distributions, of such analyses. In all cases, if the variance of the distribution is larger than its mean, the null hypothesis is rejected too frequently. Some behaviour patterns, for example, breathing, are performed regularly, i.e. the variance is less than the mean. In this case null hypotheses would be rejected too infrequently.

Frequencies of behaviour may not be independent because an individual contributes more than once to the data set. Machlis et al. (1985) discussed the problem of dependence resulting from repeated measurements on individuals. In contingency tables this could arise if one is interested in the number of interactions of each dyad in a social group. Since each individual is part of several dyads, it contributes to several cell frequencies. This affects the covariance among the cells. If one individual was particularly active socially, all cell frequencies to which this individual contributed would tend to be large. This problem of dependence can be dealt with statistically since the source of dependency is known and may be estimated (see below). A more serious problem is that the interaction rate of any dyad is unlikely to follow a Poisson distribution. If it did follow a Poisson distribution, then interactions between individuals occur randomly in time. Problems resulting from analyses where there is dependence between observations is best illustrated by example.

Table I. Five sample data sets arranged as two-by-two matrices

	Number of days									
	1		3		9		27		81	
Subject	a.m.	p.m.	a.m.	p.m.	a.m.	p.m.	a.m.	p.m.	a.m.	p.m.
1	8	11	82	38	179	156	550	558	1557	1739
2	37	20	61	74	156	196	553	609	1622	1507

Cell frequencies for 1 day were generated under the null hypothesis of independence by sampling from a discrete normal distribution of mean 19·6 and variance 95·6. Cell frequencies for the other matrices were obtained by generating cell frequencies, as described above, for 3, 9, 27 and 81 days and summing the matrices over days.

Test-of-independence Example

An ethologist wants to know whether there is independence between an individual's behaviour (subject) and time of day based on the frequency of a particular behaviour. In this example there are two subjects and two times of day (morning and evening). The null hypothesis is that an individual's behaviour and time of day are independent. If one subject performs the behaviour more often in the morning and the other more often in the evening, individual behaviour and time of day are not independent. The scientist observes how frequently the behaviour is performed by the two subjects for an hour, once in the morning and once again in the evening. Since the behaviour is not common, observations are made over several days. At the end of this time, the morning and evening frequencies are summed separately for each subject. The ethologist then constructs a two-by-two contingency table with subjects as rows and time of day as columns.

Suppose that the four cell frequencies are samples from the same distribution, i.e. the null hypothesis is true, individual behaviour and time of day are independent. Furthermore, suppose that in this distribution there are dependencies between the observations, i.e. this behaviour tends to be performed in clusters. The result is that the variance, σ^2, is greater than the mean, μ. This could arise through auto-correlation or a variety of other reasons, all yielding different probability distributions. We will model this distribution using a discrete normal distribution, where it may be easier to see effects caused by changes in variance. A discrete normal distribution is a distribution based on the continuous normal distribution where the variable, here behaviour frequencies, can only assume integer values and all probabilities of negative numbers are summed with the probability of zero behaviour patterns. For large samples, a Poisson distribution with mean λ and a discrete normal distribution with mean and variance equal to λ will be indistinguishable. Suppose that a 1-h sample of this behaviour follows a discrete normal distribution with mean 19·6 and variance 95·6. Contingency tables of simulated data coming from this distribution, summed over various numbers of days are presented in Table I.

Table II. The proportion of null hypotheses rejected out of 1000 simulated sample data sets for various *P*-values

	Proportion of rejected null hypotheses				
	Number of days				
P	1	3	9	27	81
0·5	0·793	0·769	0·789	0·758	0·753
0·1	0·537	0·463	0·505	0·477	0·444
0·05	0·450	0·365	0·424	0·398	0·355
0·01	0·322	0·238	0·273	0·260	0·225
0·001	0·213	0·123	0·155	0·144	0·133

Sample data sets were constructed identically to those in Table I. The G^2 statistic was used to test for independence.

To investigate the effects of using a G^2 test of independence on this type of data, we performed a Monte Carlo simulation using 1000 data sets. Each cell in a two-by-two contingency table was filled from a sample from the same discrete normal distribution of mean 19·6 and variance 95·6. As in the example contingency tables in Table I, the cells were summed over various numbers of days. Since

Table III. The proportion of null hypotheses rejected out of 10,000 simulated sample data sets for *P*-values of 0·05 and 0·01 from discrete normal distributions

Mean	Variance	Proportion of rejected null hypotheses	
		$P=0·05$	$P=0·01$
19·5	9·1	0·0047	0·0003
19·5	20·1	0·0569	0·0132
19·5	48·8	0·2501	0·1358
19·6	95·6	0·4353	0·1809

Sample data sets were constructed under the null hypothesis of independence from discrete normal distributions in two-by-two contingency tables. The chi-squared statistic was used to test for independence. Sample data sets represent a single day.

the cell frequencies are samples from the same distribution, one would hope that the null hypothesis, that individual behaviour (subject) and time of day are independent, would be infrequently rejected. The resulting contingency tables, analysed using the G^2 test, are given in Table II. The G^2 test rejected the null hypothesis of independence between the two factors more frequently than desired. For example, for observations summed over 9 days, 42% of null hypotheses were rejected at the 0·05 level. If the G^2 test is valid, given a true null hypothesis one expects 5% of the simulated data sets to be rejected at the 0·05 level.

Because the sum of normal distributions is a normal distribution, each cell frequency in the contingency tables of summed frequencies can be thought of as being generated from a normal distribution. The excessive rejection rate does not arise from combining observation periods, thus increasing sample size will not help. The sum of Poisson distributions is a Poisson distribution. If the true distribution of behaviour was Poisson, then the distribution of the sums would also be Poisson. The use of the G^2 statistic is invalid because the underlying assumption that the cell frequencies are samples from a Poisson distribution has been violated. We performed additional simulations with various discrete distributions using chi-squared and G^2 tests. All yielded similar results. If the mean and variance were approximately equal, the tests worked well. As the disparity between the mean and variance grew the tests became increasingly inaccurate. We give results using discrete normal distributions with similar means but different variances in Table III.

Goodness-of-fit Example

Suppose one wants to test whether behaviour patterns are performed at the same rate throughout the day, divided into 10 observation periods. Suppose that the true distribution of behaviour frequency of each observation period is identical (i.e. the null hypothesis is true), and that this distribution is a discrete uniform distribution (the probability of observing each frequency is equally likely) ranging from 0 to 14. Furthermore, suppose that observations were made for 32 days. Thus, each of the 10 observed frequencies in the contingency table consists of the sum of the frequencies of 32 observation periods made at the same time of day. The central limit theorem tells us that the distribution of this sum converges on a normal distribution. The possible range of each cell is then 0 to 448 ($=14\times32$). The expected frequency of each cell in the contingency table would be the same, calculated by summing the observed frequencies of the 10 cells and then dividing by 10.

We created 1000 simulated data sets based on the conditions described above. Using a chi-squared statistic, the hypothesis of equal frequencies was rejected 86% of the time at the 0·05 level. Since the null hypothesis is true, the test must be inappropriate. As in the test of independence, the problem arises from an inappropriate sampling distribution. The mean of this sample distribution is 224 ($=7\times32$), the variance 597 ($=18·67\times32$). The variance is greater than the mean, hence an inflated chi-squared statistic and a rejection rate considerably higher than 5%.

CONSEQUENCES

How serious are the types of errors discussed above? Unfortunately, they are probably quite serious for three reasons. First, if the data do not adhere to a Poisson distribution but are clustered, then variances are going to be greater, not smaller, than the assumed Poisson distribution. As demonstrated here, this will result in rejecting true null hypotheses more often than assumed with the selected alpha level. In one goodness-of-fit simulation using a discrete uniform distribution of 0–39 generated under the null hypothesis, 99% of the tests rejected the null hypothesis at $\alpha=0·05$. A researcher committing this error is virtually assured of rejecting a true null hypothesis. Second, increasing sample size will not aid the researcher. If the

sampling distribution of the cell entries is not Poisson, the chi-squared statistic will give similar erroneous results regardless of sample size. However, a large data set will permit one to subset the data. This may be used to help identify the shape of the original distribution. Third, intricate statistical analyses of social systems may depend heavily on the assumption of independence of behaviour (e.g. Altmann & Altmann 1977). Results using formulae developed under the assumption of independence may be misleading if behaviour is not independent. This may result in a situation where the researcher is unaware of or may not understand the assumptions of the statistical test.

We chose to use frequencies of behaviour performed by individuals to illustrate problems that can arise by using chi-squared tests. Some examples in the literature of the use of chi-squared or G^2 tests in this fashion are Emlen et al. (1975, page 158), Rissing & Pollock (1986, their Table I), and Keane (1990, page 270). However, this is by no means the only situation in which one must consider the underlying sampling distribution. Many studies, for example, Watt (1986) and Robinson (1986), use the number of encounters as the total sample size for the contingency table. The total number of encounters is then partitioned to create the cells. Under the assumption of independent Poisson sampling, all cell frequencies and the total number of encounters must be samples from Poisson distribution. At the very least, one should verify that overall encounter rate follows a Poisson distribution. Our purpose is not to give a list of every conceivable way frequency counts can violate the assumptions of statistical tests used to analyse contingency tables. If it is hard to imagine that the events being counted in each cell occur at random or that each individual in the study is completely independent of all other individuals, one should suspect that the sampling distribution is not Poisson. The burden of proof that the correct statistical procedures have been performed rests with the researcher. Unless raw data are reported in full, insufficient information is available for the reader to evaluate the correctness of a statistical analysis. This is especially true in contingency table analysis where only total cell frequencies are reported. We hope to stimulate researchers to consider the underlying distribution of their data if they plan to use chi-squared or related tests. If one suspects that cell counts do not arise from a Poisson distribution, methods given in the following section may help.

A goodness-of-fit test can be used to determine whether a distribution is Poisson. However, large sample sizes are required for reasonable power because one is concerned about type II error, incorrectly concluding that the distribution is Poisson. A graphical technique, similar to a normal probability plot, can also help identify non-Poisson distributions (Hoaglin 1980).

DEALING WITH DEPENDENT OBSERVATIONS

Simulations using several discrete distributions established that problems arose when the variance was not equal to the mean. When they were approximately equal, as they are in the Poisson distribution, both the chi-squared and G^2 tests performed well. If the variance was larger than the mean, the tests rejected the null hypothesis too often; if the variance was smaller than the mean, the tests rejected too infrequently.

If the usual contingency table analysis is inappropriate, we recommend the use of analysis of variance (ANOVA) techniques. To use them, one must satisy ANOVA assumptions and obtain a reliable estimate of within-cell variation, which can be done by subsetting the data. For example, if data are collected over a 4-h time period, a subset could be created from each 1-h segment, for a total of four subsets. As a rule of thumb, one might strive to have as many subsets as possible subject to the constraint that the average subset count is five or more. Analysis of variance assumptions may be more easily satisfied than those of the chi-squared test. In particular, independence of counts is not required. The requirement of independence for ANOVA is that the subset frequencies in each cell are independent, i.e. an observation is not counted in more than one subset. Another important assumption of ANOVA is that within-cell variances are homogeneous. Since there is often a positive relationship between means and variances in behavioural data, we suggest verifying this assumption before proceeding with parametric ANOVA.

If the sample frequencies do not appear to come from a normal distribution, the mean count of several observation periods can become the unit of analysis. Sample means taken from any distribution converge on a normal distribution (the central limit theorem). If the parent distribution is unimodal and not excessively skewed, convergence

Table IV. The proportion of null hypotheses rejected out of 1000 simulated sample data sets for *P*-values of 0·05 from various discrete distributions with known variances

			Proportion of rejected null hypotheses		
			Number of days		
Distribution	Mean	Variance	1	9	81
Discrete normal	19·6	95·6	0·055	0·051	0·050
Discrete exponential	4·7	91·6	0·007	0·032	0·047
Discrete uniform	7·1	17·8	0·053	0·053	0·046
Linearly increasing	5·2	5·4	0·067	0·046	0·042
Linearly increasing	9·8	16·0	0·050	0·063	0·038
Trimodal	5·5	9·6	0·052	0·053	0·051

Sample data sets were constructed under the null hypothesis of independence from various discrete distributions in two-by-two contingency tables. A generalized chi-squared statistic was used to test for independence.

is rapid. For example, the cumulative distribution function of a transformation of the sum of three independent uniform variables, each over the interval -1 to $+1$, differs from that of the standard normal by less than 1% (Johnson & Kotz 1970). An upper bound for the magnitude of the difference between the cumulative distribution functions of a standardized sum of N independent, identically distributed random variables and the standard normal distribution is given by Johnson & Kotz (1970). Using means, one then has greater confidence that the assumption of normality, necessary for ANOVA, has been satisfied. It would be necessary to first verify that there is little effect of observation period if, for example, observations are made at the same time of day for the same length of time each day.

Another approach if data are not normal is the use of two-way non-parametric ANOVA (Bradley 1968, page 138; Hettmansperger 1984, page 194). The non-parametric ANOVA technique, first presented by Iman (1974), and expanded on by Conover & Iman (1981), holds great promise. All observations, here subset frequencies, of the entire data set are ranked. One then performs a standard ANOVA on the ranks. If the original variables are normal, there is little loss of power, i.e. results on ranks show very high agreement with results from a standard ANOVA on the original variables. For variables that are not normal, there may be a substantial increase in power by performing the analysis on ranks rather than on the original variables. This analysis is very easy to implement, since most statistical packages will rank variables in a data set and can then be made to perform standard ANOVA on the ranks. As a cautionary note, significant F-values require a somewhat different interpretation since one is no longer testing differences of means but of medians. The use of non-parametric one-way ANOVA is also possible for testing goodness-of-fit, if the null hypothesis is that the frequencies are equal. Tests based on rank are not assumption free. However, their assumptions are less stringent than those of parametric tests since they are distribution free.

More complicated alternatives to ANOVA are available by modifying the traditional chi-squared test. One is the use of the generalized chi-squared statistic, defined as

$$\sum_{i=1}^{k} \frac{(\bar{X}_i - \lambda_i)^2}{\sigma_i^2},$$

where k is the number of cells, λ_i the parameter representing the mean of cell i, and σ_i^2 the parameter representing the variance of cell i. Simulations using a generalized chi-squared statistic were performed using various discrete distribution and based on known variances. Results from these simulations (Table IV) were encouraging, only the discrete exponential distribution failed to perform satisfactorily. If summing over a large number of observation periods, the generalized chi-squared statistic performs well since the distribution will be approximately normal. The problem of using a generalized chi-squared statistic is that one must estimate the variance of each cell as well as its mean from the data. This could be done by subsetting the

Table V. The proportion of null hypotheses rejected out of 1000 simulated sample data sets for *P*-values of 0·05 from various distributions with variances calculated from each sample

			Proportion of rejected null hypotheses				
			Number of days				
Distribution	Mean	Variance	2	4	8	16	32
Discrete uniform	5·5	11·9	0·317	0·165	0·082	0·069	0·053
			0·859	0·542	0·224	0·104	0·079
Linearly increasing	7·2	9·3	0·329	0·214	0·087	0·068	0·053
			0·824	0·574	0·243	0·126	0·075
Discrete normal	19·6	95·6	0·424	0·176	0·094	0·089	0·056
			0·913	0·529	0·211	0·105	0·074
Trimodal	5·5	9·6	0·318	0·188	0·079	0·071	0·056
			0·847	0·581	0·225	0·097	0·075

Sample data sets were constructed under the null hypothesis of independence from various discrete distributions in two-by-two (first row) and four-by-four (second row) contingency tables. A generalized chi-squared statistic was used to test for independence. The variances for the generalized chi-squared statistic were calculated from the data for each cell separately. The number of samples used to calculate the variance of each cell is the number of days over which data were summed. The variance estimate was then multiplied by the number of days for use in the generalized chi-squared statistic.

data set, as in ANOVA, or using resampling techniques, such as bootstrapping (Efron & Tibshirani 1986). In general, substantial data are necessary to reliably estimate the variance of each cell. Fewer data are needed in ANOVA because of its additional assumption of equal variance within all cells. In theory, the distribution of the generalized chi-squared statistic under the null hypothesis converges on the χ^2 distribution as the number of subsets approaches infinity. This is supported by simulations we performed using two-by-two and four-by-four contingency tables (Table V). In these simulations, the mean frequency per cell or distribution shape had less influence than the number of subsets. While the cell variance and mean must be equal in independent Poisson sampling, this does not hold for all sampling distributions yielding correct chi-squared statistics. For example, in contingency tables generated from a multinomial distribution, the theoretical cell variance is smaller than its mean. This is due to a restriction on the entire contingency table. If each cell is examined separately, however, it should hold a sample from a Poisson distribution. The mean and variance of a cell should be approximately equal if calculated using only that cell's data.

Adjustments to the chi-squared statistic can be made if it is known how and where the data depart from the assumptions of traditional chi-squared statistics. This may be useful if individuals contribute more than once to the data set, some individuals are relatives or mates, or there is some other known reason for violation of the assumption of independence. The simplest adjustment, proposed by Tideman (1979), requires an estimate of covariance between observations. He advocates using a generalized chi-squared statistic where the variance, in the denominator, is calculated by summing average variances of observations and average covariances of pairs of observations. In some data sets with repeated measures, these can be readily estimated from the data. In data where estimates of variances and covariances are impractical but the sampling design clearly identifies which observations are correlated, several authors have proposed adjusting the chi-squared statistic by dividing it with a correction factor, for example, Cohen (1976) and Rao & Scott (1987). More information is necessary to implement these techniques than will usually be available to ethologists.

Seaman & Jaeger's (1990) article on the use of parametric and non-parametric tests in ecology contains much relevant information to this discussion. In particular, they emphasize that individuals are unlikely to be distributed independently in space. Thus, effects of dependency between observations in a contingency table pertain to ecological as well as ethological data.

CONCLUSION

Our study has demonstrated some of the weaknesses of chi-squared and related tests and revealed possible consequences of their indiscriminate application to ethological data. We have shown that these tests are sensitive to the most common violation of their assumptions, lack of complete independence between observations. One should verify that the mean and variance are approximately equal to each other in each cell of a contingency table before proceeding with chi-squared and related tests. We recognize that even the best conceived study may yield insufficient data for this. However, it does no one good to report and discuss results of invalid statistical tests.

ACKNOWLEDGMENTS

This paper is partial fulfilment of the senior author's Master's degree in statistics at the University of Tennessee, Knoxville. We thank Enrique Font, Gordon Burghardt and David Sylwester for their careful reading of earlier drafts. We also thank two anonymous referees, the past assistant editor, and the past editor for constructive criticism.

REFERENCES

Altmann, S. A. & Altmann, J. 1977. On the analysis of rates of behaviour. *Anim. Behav.*, **25**, 364–372.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Massachusetts: MIT Press.

Bradley, J. V. 1968. *Distribution-free Statistical Tests*. Englewood Cliffs, New Jersey: Prentice-Hall.

Cochran, W. G. 1952. The χ^2 test of goodness of fit. *Ann. math. Stat.*, **23**, 315–345.

Cohen, J. E. 1976. The distribution of the chi-squared statistic under clustered sampling from contingency tables. *J. Am. Stat. Assoc.*, **71**, 665–669.

Conover, W. J. & Iman, R. L. 1981. Rank transformations as a bridge between parametric and nonparametric statistics. *Am. Stat.*, **35**, 124–129.

Efron, B. & Tibshirani, R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.*, **1**, 54–77.

Emlen, S. T., Rising, R. D. & Thompson, W. L. 1975. A behavioural and morphological study of sympatry in the indigo and lazuli buntings of the Great Plains. *Wilson Bull.*, **87**, 145–179.

Gleser, L. J. & Moore, D. S. 1983. The effect of dependence on chi-squared and empiric distribution tests of fit. *Ann. Stat.*, **11**, 1100–1108.

Hettmansperger, T. P. 1984. *Statistical Inference Based on Ranks*. New York: John Wiley.

Hoaglin, D. C. 1980. A Poissonness plot. *Am. Stat.*, **34**, 146–149.

Iman, R. L. 1974. A power study of a rank transform for the two-way classification model when interaction may be present. *Can. J. Stat.*, **2**, 227–239.

Johnson, N. L. & Kotz, S. 1970. *Continuous Univariate Distributions—1*. New York: John Wiley.

Lewis, D. & Burke, C. J. 1949. The use and misuse of the chi-square test. *Psychol. Bull.*, **46**, 433–489.

Keane, B. 1990. The effect of relatedness on reproductive success and mate choice in the white-footed mouse, *Peromyscus leucopus*. *Anim. Behav.*, **39**, 264–243.

Machlis, L., Dodd, P. W. D. & Fentress, J. C. 1985. The pooling fallacy: problems arising when individuals contribute more than one observation to the data set. *Z. Tierpsychol.*, **68**, 201–214.

Moore, D. S. 1986. Tests of chi-squared type. In: *Goodness-of-fit Techniques* (Ed. by R. B. D'Agostino & M. A. Stephens), pp. 63–95. New York: Marcel Dekker.

Rao, J. N. K. & Scott, A. J. 1987. On simple adjustments to chi-square tests with sample survey data. *Ann. Stat.*, **15**, 385–397.

Rissing, S. W. & Pollock, G. B. 1986. Social interaction among pleometrotic queens of *Veromessor pergandei* (Hymenoptera: Formicidae) during colony foundation. *Anim. Behav.*, **34**, 226–233,

Robinson, S. K. 1986. Competitive and mutualistic interactions among females in a neotropical oriole. *Anim. Behav.*, **34**, 113–122.

Seaman, J. W., Jr & Jaeger, R. G. 1990. Statisticae dogmaticae: a critical essay on statistical practice in ecology. *Herpetologica*, **46**, 337–346.

Sokal, R. R. & Rohlf, F. J. 1981. *Biometry*. 2nd edn. New York: W. H. Freeman.

Tideman, T. N. 1979. A generalized χ^2 for the significance of differences in repeated, related measures applied to different samples. *Educ. Psychol. Meas.*, **39**, 333–336.

Watt, D. J. 1986. A comparative study of status signalling in sparrows (genus *Zonotrichia*). *Anim. Behav.*, **34**, 1–15.